DOCUMENT RESUME

ED 385 575 TM 024 013

AUTHOR Stricker, Lawrence J.; And Others
TITLE Adjusting College Grade-Point Average for Variations in Grading Standards.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-65
PUB DATE Nov 92
NOTE 42p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Class Rank; *College Freshmen; Comparative Analysis; Correlation; *Grade Point Average; *Grading; Higher Education; *Prediction; Scores; Sex Differences; Standards; *Statistical Analysis
IDENTIFIERS *Scholastic Aptitude Test

ABSTRACT

This study compared the effectiveness of several existing and proposed methods for statistically adjusting college grade point averages (GPAs) for course and departmental differences in grading standards, using first-semester grades from an entire entering class at a large state university (4,351 students), in 1988. Most of the adjusted GPAs produced by these methods functioned similarly and, despite high correlations with actual GPA, had greater internal-consistency reliability than actual GPA and were more predictable from Scholastic Aptitude Test (SAT) scores and high school rank (HSR). Most of the adjusted GPAs also functioned similarly with regard to sex differences in over-underprediction. The adjusted GPAs and actual GPAs exhibited the same small but significant sex differences in over-underprediction by SAT scores, but the adjusted GPAs displayed smaller differences than actual GPAs in over-underprediction by SAT scores and HSR. Seven tables present analysis results. (Contains 44 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-92-65

ED 385 575

RESEARCH REPORT

ADJUSTING COLLEGE GRADE-POINT AVERAGE FOR VARIATIONS IN GRADING STANDARDS

Lawrence J. Stricker
Donald A. Rock
Nancy W. Burton
Eiji Muraki
Thomas J. Jirele

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Educational Testing Service
Princeton, New Jersey
November 1992

TM024013

BEST COPY AVAILABLE

Adjusting College Grade-Point Average for Variations in Grading Standards

Lawrence J. Stricker, Donald A. Rock, Nancy W. Burton,
Eiji Muraki, and Thomas J. Jirele

Copyright © 1992. Educational Testing Service. All rights reserved.

Abstract

This study compared the effectiveness of several existing and proposed methods for statistically adjusting college GPAs for course and departmental differences in grading standards, using first-semester grades from an entire entering class at a large state university. Most of the adjusted GPAs produced by these methods functioned similarly and, despite high correlations with actual GPA, had greater internal-consistency reliability than actual GPA and were more predictable from SAT scores and high school rank (HSR). Most of the adjusted GPAs also functioned similarly with regard to sex differences in over-underprediction. The adjusted GPAs and actual GPA exhibited the same small but significant sex differences in over-underprediction by SAT scores, but the adjusted GPAs displayed smaller differences than actual GPA in over-underprediction by SAT scores and HSR.

Adjusting College Grade-Point Average for Variations in Grading Standards

College grade-point average (GPA), though originally intended for administrative purposes (Smallwood, 1935), is widely employed in educational and psychological research, particularly as a criterion for validating admissions measures (e.g., see the reviews by Breland, 1981; Fishman & Pasanella, 1960; Lavin, 1965).

Despite the popularity of GPA, it is generally recognized that this is a fallible index of academic performance (e.g., see the reviews by Milton, Pollio, & Eison, 1986; Warren, 1971; Willingham, 1990). A major problem is that GPA is based on a different set of courses for each student, and the grading standards are not uniform from course to course, a phenomenon that has been observed for many years (e.g., Meyer, 1908). Hence, GPA is not comparable for students who take courses with severe grading standards and students who take courses with lenient standards, and its reliability and validity are attenuated.

Differences in grading standards have been rigorously documented among departments (Anderhalter, 1962; de Nevers, 1984; Elliott & Strenta, 1988; Frisbee, 1984: Gamson, 1967; Goldman & Hewitt, 1975; Goldman, Schmidt, Hewitt, & Fisher, 1974; Goldman & Widawski, 1976; Juola, 1968; Prather & Smith, 1976; Prather, Smith, & Kodras, 1979; Ramist, Lewis, & McCamley, 1990; Sabot & Wakeman-Linn, 1991; Strenta & Elliott, 1987; Willingham, 1985), as well as within departments (Garrison, 1979; Juola, 1968).

The consequences of variations in grading standards on the reliability and validity of GPA are suggested by studies that attempted to adjust GPA for differences in these standards. The adjustments increased the median correlation between yearly GPAs from .67 to .72 (Elliott & Strenta, 1988). The adjustments also generally boosted the correlations of admissions measures with GPA: the multiple correlation of the Scholastic Aptitude Test (SAT;

Donlon, 1984) scores and high school GPA with four-year GPA increased from .58 to .64 (Young, 1990b), and the correlations of the total SAT score (combining the Verbal [V] and Mathematical [M] scores) with four-year GPA went from .43 to .50 (Strenta & Elliott, 1987), but the multiple correlation of SAT scores and HSR with first-semester GPA increased from only .42 to .44 (Stricker, Rock, & Burton, 1991).

Several statistical methods that directly or indirectly adjust GPAs for differences in grading standards have been developed in recent years. Goldman and Widawski (1976) devised a within-subjects procedure that compared average course grades earned by students who took courses in different pairs of departments and then adjusted grades for the difference in these averages. (This procedure was subsequently used by Strenta and Elliott, 1987.)

Elliott and Strenta (1988) extended the Goldman and Widawski (1976) procedure, not only comparing corresponding average course grades earned by students who took courses in different pairs of departments but also comparing corresponding average course grades for students who took different courses in the same departments and adjusting grades for the differences in both averages.

Young (1990a, 1990b) applied item response theory (IRT) methods to course grades, treating the grades like polytomously scored item responses (Muraki, 1990), to secure estimated "thetas" (scores on the latent trait underlying the grades) for three fields (humanities, social sciences, and natural sciences and engineering) and then combined the three estimated thetas into a composite measure.

Stricker et al. (1991) employed a regression procedure to residualize average course grades for the characteristics of the students enrolled in the courses (high school honors courses taken in various fields, intended college

majors, percentage of college-bound seniors in their high schools) and adjusted grades for the residual.

Other procedures are also applicable to this problem. The discrepancy between the average grade in a course and the average predicted overall GPA (predicted from admissions test scores and high school record) for students in the course could be used to adjust course grades. Such an index (with SAT scores and high school GPA as predictors), the "grade-residual mean," was used recently by Ramist et al. (1990) to assess the leniency or severity of the grading standards for courses. Variants of this index were employed for the same general purpose in previous studies. Anderhalter (1962) used the discrepancy between the average grade for a department and average predicted overall GPA (based on admissions test scores and HSR) to evaluate departments' grading standards. And Juola (1968) employed the difference between the average grade for a course and average actual overall GPA (in other courses) to assess courses' grading standards.

The unavailable grades for courses that students do not take could be treated as a missing data problem (Little & Rubin, 1987), with the missing grades considered as "missing at random" in the sense that they are predictable from available grades. The missing grades can be imputed by maximum likelihood methods, using the EM algorithm (Dempster, Laird, & Rubin, 1977), generating a complete set of grades for all students.

The Stricker et al. (1991) regression method could be modified to eliminate student characteristics that are specific to disciplines and hence may undercorrect for departmental differences in grading standards.

These existing and proposed methods for adjusting GPA differ in whether they rely on "internal" data (other grades) or "external" data (other, non-grade variables); the methods also vary in their complexity and

sophistication. But nothing is known about the methods' relative effectiveness in improving the psychometric properties of GPA. Accordingly, the main aim of the present study was to compare these methods with regard to their intercorrelations, reliability, and correlations with admissions measures. A secondary purpose was to assess the effects of these methods on sex differences in over-underprediction, for several of the methods have been applied to this problem (Elliott & Strenta, 1988; Stricker et al., 1991; Young, 1991). A final goal was to explore the efficacy of a novel approach, suggested by Ramist et al. (1990), for predicting GPA from the cumulated predictions of individual course grades.

Method

Sample

The sample consisted of 4,351 students (2,318 women and 2,033 men) in the Fall 1988 entering class at a large state university's main campus. The sample was limited to full-time freshmen enrolled in the seven undergraduate schools: three liberal arts colleges and four professional schools. This is the same sample used in the Stricker et al. (1991) study.

Actual and Adjusted GPAs

First-semester grades in all degree-credit courses were used. (No Credit grades assigned to students in one of the liberal arts colleges in lieu of Fs were treated as Fs, and temporary grades were treated the same as permanent ones.) The cohort enrolled in 498 courses in 86 departments. These grades were also used in the Stricker et al. (1991) study.

Whenever possible, adjusted GPAs were based on the 140 individual and pooled courses used with the original Stricker et al. (1991) procedure, described subsequently, which provides a means of adjusting grades for all courses, regardless of their size. (The IRT GPA, as used originally by Young,

1990a, 1990b, was restricted to courses with a minimum size.) The exceptions were the within-subjects GPA, which had provisions for dealing with the size problem, and the imputed GPA, which was computationally impractical to apply to the 140 individual and pooled courses. The actual and adjusted GPAs derived from the grades were based on all grades and weighted by the number of credit hours per course, unless otherwise noted.

Actual GPA. This GPA is based on actual, unadjusted grades; it is the same variable used in the Stricker et al. (1991) study.

Within-subjects GPA. The Elliott and Strenta (1988) procedure was followed with the 498 individual courses. Between-department adjustments were made for 53 individual departments and a pooled department that combined departments with fewer than ten grades. (Between-department discrepancies were weighted by the number of students involved.) Within-department adjustments were made in 12 departments for 33 individual courses and a pooled course that combined courses in the same department that had fewer than ten grades.

IRT GPA. The Young (1990a, 1990b) procedure was followed with the same 140 individual and pooled courses used in the original Stricker et al. (1991) method. These are 119 individual courses with available data for ten or more students and 21 pooled courses that combine individual courses to achieve sample sizes of ten or more students: 20 pooled courses made up of individual courses combined by department and one pooled course comprised of individual courses combined across departments. Courses were categorized as humanities, social sciences, natural sciences, or other, using an adaptation of the university's department classification employed in the Stricker et al. (1991) study. The IRT analyses, employing the PARSCALE program (Muraki & Bock, 1991), were done separately for the major categories of courses: 55

humanities courses (N = 3,874), 32 social sciences courses (N = 2,639), and 44 natural sciences courses (N = 3,342). (Two humanities courses with no variation in grades, four "other" courses, and the pooled course that combined individual courses across departments were excluded.) The estimated thetas for the three fields were then standardized, using the data for the 1,651 students with all three estimated thetas. An analog to GPA was computed, weighting each standardized estimated theta by the number of credit hours in the same course category. (The composite measure used by Young, 1990b, 1991, was computed differently: it is the weighted average of the unstandardized estimated thetas, each estimated theta weighted by the square root of the reciprocal of its standard error of estimate.)

Imputed GPA. Maximum likelihood estimates of the GPAs in each of 53 individual departments and a pooled department that combined departments with fewer than ten GPAs were obtained, with the BMDP AM Program (Frane, 1990) using available GPAs for the 54 departments. (Using department GPAs instead of course grades facilitated estimation by reducing the size and sparseness of the student-by-grade data matrix.) An unweighted overall GPA was computed. (GPA was unweighted because of the unavailability of the number of credit hours per department.)

Original regression GPA. This GPA is based on the original Stricker et al. (1991) procedure, applied to 140 individual and pooled courses. This is the same variable used in the Stricker et al. (1991) study.

Modified regression GPA. The original Stricker et al. (1991) procedure was modified by changing some of the variables that describe the students in each course. Three variables were employed: Percentage with High School Honors Courses in Any Field,[1] Percentage of College-Bound Seniors in their High School, and Percentage with Data on High School Honors Courses. The

source of the data on honors courses was the Student Descriptive Questionnaire, completed by students when they registered for the SAT, and recorded in the university's Longitudinal Data Base (LDB). The source of the data on college-bound seniors was the Attending Institution Profile Survey of high school officials conducted by Educational Testing Service in 1988; the student's high schools were recorded in the LDB. Percentage with Data on High School Honors Courses was included to adjust for the effect of missing data on Percentage with High School Honors Courses in Any Field by capitalizing on the information inherent in the presence or absence of data for the latter variable (J. Cohen & P. Cohen, 1983). (Data on this variable was missing for 25.0% of the sample; data on college-bound seniors was missing for only 1.4%.) The same 140 individual and pooled courses used with the original Stricker et al. (1991) procedure were employed.

Grade-residual GPA. The Ramist et al. (1990) method was followed with the individual and pooled courses used in the original Stricker et al. (1991) procedure. (HSR was substituted for the unavailable high school GPA used by Ramist et al., 1990.) The difference between mean course grade and mean predicted overall GPA was then applied to the grades of each student in the course, including those without predicted GPAs.

Predicted GPA. The predicted GPA proposed by Ramist et al. (1990) was calculated, following their method for predicting individual course grades and using the 140 courses in the original Stricker et al. (1991) procedure. (HSR was substituted for the high school GPA used by Ramist et al., 1990.) For each course, a regression equation of SAT-V, SAT-M, and HSR against course grade was calculated (deleting predictors with negative correlations with course grade or negative regression coefficients), and predicted course grades

were obtained with the equation. A GPA was computed from these predicted grades.

Other Variables

Sex, SAT scores, and HSR were obtained from the LDB; HSR was converted to normalized T scores. The original source of these variables, also used in the Stricker et al. (1991) study, was official records.

Analyses

Analyses were conducted for the actual and adjusted GPAs, and for two kinds of residualized actual and adjusted GPAs that represented over-underprediction. One kind of residualized GPA used predictions from SAT-V and SAT-M; the other kind used predictions from SAT-V, SAT-M, and HSR. The predictions were made with regression equations for students in the cohort with complete data on the particular set of predictors (SAT scores or SAT scores and HSR) and GPA. (The same analyses of over-underprediction were conducted in the Stricker et al., 1991, study.)

Similar analyses were done for the predicted GPA measure and an analogous measure of over-underprediction: actual GPA residualized for predicted GPA. Because predicted GPA is derived from SAT scores and HSR, the corresponding residualized GPA measure is included only in analyses of GPAs residualized for both kinds of predictors.

Product-moment intercorrelation matrices were computed, using missing data procedures (each correlation was based on all available students), and multiple correlations were calculated from these matrices.

Because of the large sample size, both statistical and practical significance were considered in assessing the results. The .01 level was used throughout in view of the sample size. (The total N for significance tests of multiple correlations was the smallest N for any of the zero-order

correlations involved.) A minimum effect size (J. Cohen, 1988) was used that accounted for 1% of the variance (e.g., a correlation of .10, a difference in means [d] of .20 of a standard deviation.) This size is commonly considered to be a "small" effect from the standpoint of practical significance (J. Cohen, 1988).

The internal-consistency reliabilities of the actual and adjusted GPAs were estimated. For the actual GPA, within-subject GPA, original regression GPA, modified regression GPA, and grade-residual GPA, GPAs for "odd" and "even" halves of the course grades were obtained, and reliability was estimated by the Spearman-Brown formula from the correlations between the two GPAs. For the IRT GPA, GPAs for each of three fields were obtained, weighting each standarized theta by the corresponding number of credit hours; reliability was estimated by the Spearman-Brown formula from the mean intercorrelation between the three GPAs. And for the imputed GPA, reliability was estimated by Coefficient Alpha.

Results

Intercorrelations of Actual and Adjusted GPAs

The intercorrelations and internal-consistency reliabilities for the actual and adjusted GPAs appear in Table 1. All the GPAs, including actual GPA, correlated highly with each other (.91 to .99), but the IRT GPA consistently correlated lower than the others (.91 to .94).

The GPAs' reliabilities varied considerably (.64 to .99). The imputed GPA (.99), grade residual GPA (.77), original regression GPA (.76), and modified regression GPA (.76) had higher reliabilities than actual GPA (.70), and the IRT GPA (.64) had a lower reliability; the reliability of the within-subjects GPA (.71) was similar to that of actual GPA.

Insert Table 1 about here

Correlations of SAT Scores and HSR with Actual and Adjusted GPAs

The zero-order and multiple correlations of SAT scores and HSR with actual and adjusted GPAs appear in Table 2. The SAT scores and HSR generally correlated higher with the adjusted GPAs than with the actual GPA. The original Stricker et al. GPA was an exception: the correlations of SAT scores and HSR with it were close to those with actual GPA. The correlations with the other adjusted GPAs were generally similar.

The predicted GPA correlated .56 ($p < .01$) with actual GPA, somewhat larger than the corresponding multiple correlation of .42 for SAT scores and HSR with this criterion. Note that the former correlation is inflated because actual grades in individual courses are used as criteria in the process of obtaining predicted grades for these courses, and these same actual course grades, in turn, are the components of actual GPA.

Insert Table 2 about here

Intercorrelations of GPA Over-Underprediction Measures

The intercorrelations of the actual and adjusted GPAs residualized for SAT appear in Table 3. Paralleling the intercorrelations of actual and adjusted GPAs, all these residualized GPAs correlated highly with each other (.90 to .99), but the IRT measure consistently correlated lower with the others (.90 to .93).

Insert Table 3 about here

The corresponding intercorrelations of the actual and adjusted GPAs residualized for SAT and HSR are shown in Table 4. The predicted GPA residualized measure also appears in this table. All the residualized GPAs, including the predicted GPA measure, correlated highly (.88 to .99). But the IRT measure (.88 to .92) and the predicted GPA measure (.88 to .96) correlated lower than the others.

Insert Table 4 about here

Sex Differences in GPA Over-Underprediction Measures

The mean actual and adjusted GPAs residualized for SAT scores are reported in Table 5 for women and men; the corresponding statistics for the GPAs residualized for SAT scores and HSR appear in Table 6, together with the statistics for the predicted GPA measure. For comparison, the mean actual and adjusted GPAs for both sexes appear in Table 7. Note that the *d* indexes for differences between the means for women and men and for the statistical significance of these differences are inflated for imputed GPA and the imputed GPA residualized measures because the variability of imputed GPA is attenuated by the imputation process (Little and Rubin, 1987); the actual differences between the means for these variables are unaffected. In addition, the actual differences between the means for IRT GPA and the IRT GPA residualized measures are not comparable to those for actual GPA and other adjusted GPAs because IRT GPA is not on the same 1-4 grade scale; the *d* indexes for these IRT GPA variables are comparable. And the differences between the means for the predicted GPA residualized measure are underestimated for the reasons mentioned previously.

All the sex differences for actual and adjusted GPAs (Table 7) were small (actual differences of .01 to .09; ds of .00 to .11), and most were not significant ($p > .01$), with the exception of actual GPA and the original regression GPA. All the adjusted GPAs, except the original regression GPA, had smaller sex differences than actual GPA.

All the sex differences for actual and adjusted GPAs residualized for SAT (Table 5) were small (actual differences of -.09 to -.21; ds of -.21 to -.26) but statistically significant ($p < .01$). The sex differences were generally similar for the actual and adjusted GPA measures, but were somewhat smaller for the imputed GPA measure.

All the corresponding differences for actual and adjusted GPAs residualized for SAT and HSR (Table 6) were substantially smaller than those for GPAs residualized for SAT. The sex differences were small (actual differences of -.04 to -.11; ds of -.05 to -.15) and, except for the predicted GPA measure, were statistically significant ($p < .01$). All the adjusted GPA measures, except the original regression measure, displayed substantially smaller sex differences than the actual GPA measure. The other adjusted GPA measures generally had similar sex differences, but the differences for the imputed GPA and predicted GPA measures were somewhat smaller.

Insert Tables 5 to 7 about here

Discussion

Psychometric Properties of Adjusted GPAs

A central finding is that most of the methods for adjusting GPA, with the exception of the original regression procedure, functioned similarly and, despite high correlations with actual GPA, operated differently from it. The

adjusted GPAs generally appeared to be psychometrically superior in reliability and, on the basis of their predictability from SAT and HSR, in validity. The evidence on the latter point is only suggestive and needs to be confirmed by further investigations of the comparative validity of adjusted and actual GPA, for it is at least conceivable that the enhanced predictability of the adjusted GPAs could come about for reasons extraneous to academic success (e.g., the common effects of test anxiety on both the admissions measures and the adjusted grades) that are unintentionally magnified by the adjustment process.

It should be recognized that the adjustment methods are not free from problems. Most of the methods, with the exception of the original regression and imputed procedures, directly or indirectly emphasize general ability, and hence may make inadequate adjustments for grades in courses that demand special abilities or interests, such as courses in the arts, or involve unusually superior or inferior instruction (Strenta & Elliott, 1987). Insofar as there are many such courses, the validity of the adjusted GPA will be affected.

In addition, though actual GPA is far from perfect from a psychometric perspective, its flaws should not be overstated (Etaugh, Etaugh, & Hurd, 1972). Actual GPA's reliability is substantial, its predictability is appreciable, and it is factorially simple. (Schoenfeldt and Brush, 1975, found a large general factor and a smaller agriculture and education factor in an analysis of cumulative GPAs over 13 quarters in 12 fields; Young, 1990a, 1990b, identified two group factors, natural sciences and engineering, and social sciences and humanities, in an analysis of freshmen grades in 127 courses.

All the methods, with the possible exception of the original regression procedure, appear promising, given the limited data currently available about them. Of these methods, the grade residual procedure is the most desirable from the standpoint of computational simplicity.

The limited effectiveness of both the original regression method and the new regression method in adjusting grades probably stems from their use of variables that are only indirectly and weakly related to college grades. (The multiple correlations of the variables with the mean course grades were .47 for the original regression method and .31 for the new regression method.) All the other methods, in contrast, relied on either college grades or SAT scores and HSR, much more potent variables.

Incidentally, the similarity in functioning of the original regression method, which used major-specific variables, and the new regression method, which did not, indicates that undercorrecting grades for department differences was not the explanation for the original regression method's minimal success in adjusting grades.

The present results are generally similar to those obtained in previous studies. The reliability for actual GPA is in the same range as the internal-consistency estimates for freshman GPA observed earlier (Barritt, 1966; Clark, 1964; Etaugh et al., 1972; Millman, Slovacek, Kulick, & Mitchell, 1983; Ramist et al., 1990; Singleton & Smith, 1978). But the failure of the within-subjects method to enhance reliability appears inconsistent with the higher correlations, observed by Elliott and Strenta (1988), between academic-year GPAs based on this method.

The high correlation between the IRT GPA and actual GPA resembles the equally high correlations, reported by Young (1991), between GPA adjusted by this method and actual GPA (both GPAs for four academic years).

The higher correlations of SAT-V and SAT-M with the within-subjects GPA than with the actual GPA are consistent with the higher correlations of total SAT scores with the four-year GPA adjusted by this method that Elliott and Strenta (1988) reported. And the higher correlations of SAT scores and HSR with the IRT GPA than with actual GPA are comparable to the higher correlations of SAT scores and high school GPA with four-year GPA adjusted by this method, observed by Young (1990b; 1991).

Future efforts at adjusting grades might benefit from combining features of the internal and external methods. One obvious approach is to obtain adjusted grades with an internal method (within-subjects, IRT, or imputed), and then modify these grades with an external method (original regression, modified regression, or grade residual). An alternative is to incorporate auxiliary information about the courses or the students in making IRT estimates of course performance (Embretson [Whitely], 1984; Mislevy, 1987) or in imputing course grades. This research might also profit from using a broad range of course variables (including characteristics of their instructors and the students enrolled in them) to adjust grades (Frisbee, 1984; Prather and Smith, 1976).

Sex Differences in GPA Over-Underprediction

Another important outcome was that the various adjusted GPAs generally functioned similarly with regard to sex differences in over-underprediction. (The original regression method was an exception.) The methods reduced or even eliminated differences in GPA. However, they failed to narrow over-underprediction by SAT scores, though they did cut over-underprediction by SAT scores and HSR. The greater effectiveness of the grade adjustments in reducing over-underprediction by SAT scores and HSR is intriguing but cannot be explained at this point. It is evident, though, that the grade adjustments

and HSR operate independently to reduce the over-underprediction and reflect different processes at work in students' course selection and grade getting.

The considerably smaller amount of over-underprediction when HSR was added to SAT scores reflects the incremental validity of HSR in predicting grades. Given the limited amount of over-underprediction in this situation, the sizable reduction produced by the adjustments is remarkable.

The small but significant sex differences in over-underprediction associated with all the adjusted GPAs clarify inconsistencies in the findings between the Stricker et al. (1991) study and other investigations. In the Stricker et al. investigation, modest but statistically significant sex differences in over-underprediction were found with the original regression procedure, in contrast with insignificant sex differences in over-underprediction observed with the within-subjects and the IRT methods in studies at other universities (Elliott & Strenta, 1988; Young, 1991). The uniformly significant sex differences in over-underprediction in the present study, regardless of the adjustment method used, reinforce the Stricker et al. contention that the discrepant outcomes in their investigation and in the previous studies are not attributable to the various methods employed, but probably reflect institutional differences.

Predicted GPA Measure

The functioning of the predicted GPA measure was striking. It had appreciably greater effectiveness than the SAT scores and HSR in predicting actual GPA, though it was based on these measures, and exhibited no significant sex differences in over-underprediction in contrast to the significant, though small, differences in over-underprediction displayed by the SAT and HSR.

These estimates of the effectiveness of the predicted GPA measure are inflated by capitalization on chance, but they still indicate this measure's potential. The measure obviously needs to be cross validated to obtain precise estimates of its effectiveness. The present equations for course grade predictions could be applied to the same courses in a subsequent year; alternatively, new prediction equations could be obtained, using half of the study cohort, and applied to the other half.

From the standpoint of improving the prediction of academic performance, it remains to be seen whether this method will prove to be as effective in dealing with differences in grading standards by modifying the predictor as grade adjustment methods are by modifying the criterion.

References

Anderhalter, O. F. (1962). Developing uniform departmental grading standards in a university. Journal of Experimental Education, 31, 210-211.

Barritt, L. S. (1966). Note: The consistency of first-semester college grade point average. Journal of Educational Measurement, 3, 261-262.

Breland, H. M. (1981). Assessing student characteristics in admissions to higher education (Research Monograph 9). New York: College Board.

Clark, E. L. (1964). Reliability of grade-point averages. Journal of Educational Research, 57, 428-430.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd. ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., & Cohen, P. (1983). Applied multiple correlation/regression analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1-22.

de Nevers, N. (1984). An engineering solution to grade inflation. Engineering Education, 74, 661-663.

Donlon, T. F. (Ed.) (1984). The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board.

Elliott, R., & Strenta, A. C. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. Journal of Educational Measurement, 25, 333-347.

Embretson (Whitely), S. (1984). A general latent trait model for response processes. Psychometrika, 49, 175-186.

Etaugh, A. F., Etaugh, C. F., & Hurd, D. E. (1972). Reliability of college grades and grade point averages: Some implications for prediction of academic performance. Educational and Psychological Measurement, 32, 1045-1050.

Fishman, J. A., & Pasanella, A. K. (1960). College admission-selection studies. Review of Educational Research, 30, 298-310.

Frane, J. (1990). Description and estimation of missing data. In W. J. Dixon (Ed.), BMDP statistical software manual. Los Angeles, CA: BMDP.

Frisbee, W. R. (1984). Course grades and academic performance by university students: A two-stage least squares analysis. Research in Higher Education, 20, 345-365.

Gamson, Z. F. (1967). Performance and personalism in student-faculty relations. Sociology of Education, 40, 279-301.

Garrison, D. A. (1979). Measuring differences in the assigning of grades. Improving College and University Teaching, 27, 68-71.

Goldman, R. D., & Hewitt, B. N. (1975). Adaptation-level as an explanation for differential standards in college grading. Journal of Educational Measurement, 12, 149-161.

Goldman, R. D., Schmidt, D. E., Hewitt, B. N., & Fisher, R. (1974). Grading practices in different major fields. American Educational Research Journal, 11, 343-357.

Goldman, R. D., & Widawski, M. H. (1976). A within-subjects technique for comparing college grading standards: Implications in the validity of the evaluation of college achievement. Educational and Psychological Measurement, 36, 381-390.

Juola, A. E. (1968). Illustrative problems in college-level grading. Personnel and Guidance Journal, 47, 29-33.

Lavin, D. E. (1965). The prediction of academic performance. New York: Russell Sage Foundation.

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley.

Meyer, M. (1908). The grading of students. Science, 28 (712), 243-250.

Millman, J., Slovacek, S. P., Kulick, E., & Mitchell, K. J. (1983). Does grade inflation affect the reliability of grades? Research in Higher Education, 19, 423-429.

Milton, O., Pollio, H. R., & Eison, J. A. (1986). Making sense of college grades. San Francisco: Jossey-Bass.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. Applied Psychological Measurement, 11, 81-91

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. Applied Psychological Measurement, 14, 59-71.

Muraki, E., & Bock, R. D. (1991). PARSCALE: Parameter scaling for rating data. Chicago, IL: Scientific Software.

Prather, J. E., & Smith, G. (1976). A study of the relationships between faculty characteristics, subject field, and course grading patterns. Research in Higher Education, 5, 351-363.

Prather, J. E., Smith, G. & Kodras, J. E. (1979). A longitudinal study of grades in 144 undergraduate courses. Research in Higher Education, 10, 11-24.

Ramist, L., Lewis, C. & McCamley, L. (1990). Implications of using freshman GPA as the criterion for the predictive validity of the SAT. In W. W. Willingham, C. Lewis, R. Morgan, & L. Ramist (Eds.), Predicting college grades: An analysis of institutional trends over two decades (pp. 253-288). Princeton, NJ: Educational Testing Service.

Sabot, R., & Wakeman-Linn, J. (1991). Grade inflation and course choice. Journal of Economic Perspectives, 5, 159-170.

Schoenfeldt, L. F., & Brush, D. H. (1975). Patterns of college grades across curricular areas: Some implications for GPA as a criterion. American Educational Research Journal, 12, 313-321.

Singleton, R., Jr., & Smith, E. R. (1978). Does grade inflation decrease the reliability of grades? Journal of Educational Measurement, 15, 37-41.

Smallwood, M. L. (1935). An historical study of examinations and grading systems in early American universities. Cambridge, MA: Harvard University Press.

Strenta, A. C., & Elliott, R. (1987). Differential grading standards revisited. Journal of Educational Measurement, 24, 281-291.

Stricker, L. J., Rock, D. A., & Burton, N. W. (1991). Sex differences in SAT predictions of college grades (College Board Report 91-2; ETS Research Report 91-38). New York: College Board.

Warren, J. R. (1971). College grading practices: An overview (Report 9). Washington, DC: ERIC Clearinghouse on Higher Education.

Willingham, W. W. (1990). Understanding yearly trends. In W. W. Willingham, C. Lewis, R. Morgan, & L. Ramist (Eds.). Predicting college grades: An analysis of institutional trends over two decades (pp. 23-84). Princeton, NJ: Educational Testing Service.

BEST COPY AVAILABLE

Willingham, W. W. (1985). Success in college. New York: College Entrance Examination Board.

Young, J. W. (1990a). Adjusting the cumulative GPA using item response theory. Journal of Educational Measurement, 27, 175-186.

Young, J. W. (1990b). Are validity coefficients understated due to correctable defects in the GPA? Research in Higher Education, 31, 319-325.

Young, J. W. (1991). Gender bias in predicting college academic performance: A new approach using item response theory. Journal of Educational Measurement, 28, 37-47.

Author Note

Thanks are due to Leonard Ramist for advising on grade adjustment methods; Min hwei Wang for computer programming; and Hunter M. Breland and Leonard Ramist for reviewing a draft of this report.

Footnote

[1]More precisely, this is the weighted percentage of those taking (or planning to take) courses in six disciplines (Arts and Music, English, Foreign and Classical Languages, Mathematics, Natural Sciences, Social Sciences and History) who were (or planned to be) in honors, advanced placement, or accelerated courses.

Table 1

Intercorrelations of Actual and Adjusted GPAs

GPA	Mean	S.D.	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Actual	2.57	.85	(.70)	.96	.91	.95	.99	.98	.96
(2) Within-subjects	2.66	.85		(.71)	.94	.96	.98	.98	.99
(3) IRT	-.06	1.02			(.64)	.92	.92	.93	.94
(4) Imputed	2.81	.44				(.99)	.95	.96	.96
(5) Original regression	2.59	.83					(.76)	.99	.98
(6) Modified regression	2.59	.82						(.76)	.99
(7) Grade residual	2.55	.84							(.77)

Note. Internal-consistency reliability estimates appear in the diagonal. Ns vary from 4,306 to 4,307. All correlations are significant at the .01 level (two-tail).

Table 2

Correlations of SAT Scores and HSR with Actual and Adjusted GPAs

	SAT Scores and HSR				
GPA	SAT-V	SAT-M	HSR	SAT-V, SAT-M[a]	SAT-V, SAT-M, HSR[b]
Actual	.33	.30	.34	.36	.42
Within subjects	.33	.38	.41	.42	.50
IRT	.32	.36	.41	.40	.49
Imputed	.32	.35	.40	.39	.47
Original regression	.32	.31	.36	.37	.44
Modified regression	.31	.35	.37	.38	.45
Grade residual	.35	.40	.41	.43	.51

Note. Ns vary from 4,267 to 4,268 for SAT-V and SAT-M zero-order correlations and for SAT-V and SAT-M multiple correlations, from 3,989 to 3,990 for HSR zero-order correlations, and from 3,965 to 3,966 for SAT-V, SAT-M, and HSR multiple correlations. All correlations are significant at the .01 level (two-tail for zero-order correlations).

[a]This is the multiple correlation for SAT-V and SAT-M.

[b]This is the multiple correlation for SAT-V, SAT-M, and HSR.

Table 3

Intercorrelations of Actual and Adjusted GPA Residualized Criteria (SAT)

GPA	Mean	S.D.	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Actual	.00	.79		.96	.90	.94	.98	.98	.96
(2) Within subjects	.00	.78			.93	.96	.98	.98	.99
(3) IRT	.00	.93				.91	.91	.92	.93
(4) Imputed	.00	.41					.95	.95	.96
(5) Original regression	.00	.77						.99	.98
(6) Modified regression	.00	.76							.99
(7) Grade residual	.00	.76							

Note. Ns vary from 4,267 to 4,268. All correlations are significant at the .01 level (two-tail).

Table 4

Intercorrelations of Actual and Adjusted GPA Residualized Criteria (SAT/HSR)

GPA	Mean	S.D.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Actual	.00	.77		.96	.90	.94	.98	.98	.96	.88
(2) Within-subjects	.00	.74			.92	.95	.97	.98	.99	.94
(3) IRT	.00	.89				.90	.91	.91	.92	.88
(4) Imputed	.00	.39					.95	.95	.95	.91
(5) Original regression	.00	.75						.99	.98	.93
(6) Modified regression	.00	.73							.99	.95
(7) Grade residual	.00	.72								.96
(8) Predicted GPA	.00	.71								

Note. Ns vary from 3,965 to 3,966. All correlations are significant at the .01 level (two-tail).

Table 5

Sex Differences in Actual and Adjusted GPA Residualized Criteria (SAT)

GPA Residualized Criterion	Women			Men			Mean Difference	
	N	Mean	S.D.	N	Mean	S.D.	Actual	d[a]
Actual	2,283	-.10	.75	1,985	.11	.82	-.21**	-.26
Within subjects	2,283	-.09	.73	1,985	.10	.81	-.19**	-.24
IRT	2,282	-.09	.86	1,985	.11	1.00	-.20**	-.21
Imputed	2,283	-.04	.39	1,985	.05	.42	-.09**	-.24
Original regression	2,283	-.09	.73	1,985	.11	.80	-.20**	-.26
Modified regression	2,283	-.08	.72	1,985	.09	.79	-.17**	-.21
Grade residual	2,283	-.08	.72	1,985	.09	.80	-.17**	-.22

[a] d (J. Cohen, 1988) = (Women Mean - Men Mean)/Total Standard Deviation.

** $p < .01$, two-tail.

Table 6

Sex Differences in Actual and Adjusted GPA Residualized Criteria (SAT/HSR)

GPA Residualized Criterion	Women			Men			Mean Difference	
	N	Mean	S.D.	N	Mean	S.D.	Actual	d[a]
Actual	2,135	-.05	.74	1,831	.06	.79	-.11**	-.15
Within subjects	2,135	-.04	.71	1,831	.04	.77	-.08**	-.11
IRT	2,134	-.03	.83	1,831	.04	.94	-.07**	-.08
Imputed	2,135	-.02	.38	1,831	.02	.40	-.04**	-.10
Original regression	2,135	-.05	.72	1,831	.06	.78	-.11**	-.15
Modified regression	2,135	-.03	.71	1,831	.04	.76	-.07**	-.10
Grade residual	2,135	-.03	.70	1,831	.03	.75	-.06**	-.09
Predicted GPA	2,135	-.02	.68	1,831	.02	.74	-.04	-.05

[a] d (J. Cohen, 1988) = (Women Mean - Men Mean)/Total Standard Deviation.

** $p < .01$, two-tail.

Table 7

Sex Differences in Actual and Adjusted GPAs

	Women			Men			Mean Difference	
GPA	N	Mean	S.D.	N	Mean	S.D.	Actual	d[a]
Actual	2,293	2.61	.81	2,014	2.52	.88	.09**	.11
Within subjects	2,293	2.67	.81	2,014	2.64	.90	.03	.03
IRT	2,292	-.05	.95	2,014	-.07	1.10	.02	.02
Imputed	2,293	2.82	.43	2,014	2.80	.46	.02	.05
Original regression	2,293	2.63	.79	2,014	2.55	.87	.08**	.10
Modified regression	2,293	2.60	.79	2,014	2.57	.86	.03	.03
Grade residual	2,293	2.55	.80	2,014	2.54	.88	.01	.00

[a] d (J. Cohen, 1988) = (Women Mean - Men Mean)/Total Standard Deviation.

** $p < .01$, two-tail.